Building an Ontology Catalogue for Smart Cities

M. Poveda-Villalón, R. García-Castro and A. Gómez-Pérez
Ontology Engineering Group, Escuela Técnica Superior de Ingenieros Informáticos
Universidad Politécnica de Madrid, Spain
{*mpoveda,rgarcia,asun*}*@fi.upm.es*

ABSTRACT: Apart from providing semantics and reasoning power to data, ontologies enable and facilitate interoperability across heterogeneous systems or environments. A good practice when developing ontologies is to reuse as much knowledge as possible in order to increase interoperability by reducing heterogeneity across models and to reduce development effort. Ontology registries, indexes and catalogues facilitate the task of finding, exploring and reusing ontologies by collecting them from different sources. This paper presents an ontology catalogue for the smart cities and related domains. This catalogue is based on curated metadata and incorporates ontology evaluation features. Such catalogue represents the first approach within this community and it would be highly useful for new ontology developments or for describing and annotating existing ontologies.

1 INTRODUCTION

Ontologies allow developers to reuse and share application domain knowledge using a common vocabulary across heterogeneous systems or environments. Therefore, ontologies do not only provide semantics and reasoning power to the data described in a given application but also increase the interoperability with other data. One good practice when developing ontologies is to reuse as much knowledge as possible since it increases interoperability by reducing heterogeneity across models and reduces development time.

There are many ontologies that can be used to describe cross-domain data or data in specific domains (e.g., smart cities, energy). Therefore, developers should reuse those models that have been already created, validated and deployed rather than spending resources reinventing the wheel.

However, since these resources are spread over the Web, it is not trivial to find them; hence, a centralized place for discovering them would be helpful for new developers willing to reuse ontologies. Nowadays, there are several search engines and ontology registries that could help finding and locating ontologies; however, they are either too general or cover concrete unrelated domains (e.g., biology). This shows the need for gathering together specific ontologies for smart cities, energy and other related domains. Finally, as the quality of ontologies in the Web is unknown, a system providing quality indicators about ontologies would help developers in order to select or assess potential ontologies to be reused.

This paper presents an ontology catalogue (available at http://smartcity.linkeddata.es/ontologies/) for the smart cities and related domains in which information has been curated and that incorporates ontology evaluation features. This catalogue is the first approach within this community and it would be highly useful for new ontology developments or for describing and annotating existing ontologies.

In addition, the ontology catalogue is published both in human-readable and machine-processable formats. More precisely, the catalogue is published as an HTML web site for humans and in the RDF format (Brickley 2004) for machines. Within the RDF representation links to other data sources have been stablished, linking entities in the catalogue to entities in other datasets, contributing therefore to the Linked Data initiative (http://www.w3.org/standards/semanticweb/data).

The structure of the paper is the following. Section 2 summarizes the relevant domains for the catalogue being developed. Section 3 introduces and describes the proposed approach to collect ontology metadata. Section 4 describes the model used to represent ontology metadata in RDF while section 5 is devoted to ontology evaluation. Sections 6 and 7 describe the catalogue generation and publication in RDF and HTML, respectively. Finally, section 8 exposes related research efforts and Section 9 presents some concluding remarks and future lines of work.

2 RELEVANT DOMAINS

As already mentioned, while there are ontology registries and indexes that are general and cover a broad range of topics, there is a gap for them in the area of smart cities. For this reason, the catalogue is focused on ontologies in domains closely related to smart cities and on cross-domain ontologies that might be useful in this field. In this sense, we are focused on the following domains from the smart cities and other close domains, such as:

1. Energy (e.g., energy type, energy demand, energy offer)
2. Climate (e.g., climate zone, rainfall, sunshine hours)
3. Weather (e.g., outside temperature, wind speed)
4. Environment (e.g., pollution)
5. Building (e.g., building characteristics, insulation, spatial location, postal address, owner, manager)
6. Occupancy (e.g., based on users schedule, etc.)
7. User behavior and characteristics (e.g., practices for using devices)

Regarding the cross-domain ontologies we consider the following domains:

1. Temporal (e.g., date, time, interval)
2. Organizational (e.g., entity, legal identity, contracts, financial standing, etc.)
3. Statistical (e.g., algorithms, statistical methods, baselines, control groups)
4. Spatial (e.g., location, coordinates)
5. Measurement (e.g., scales, metrics, units, classifications)

3 ONTOLOGY COLLECTION

Different methods are being used for collecting ontologies. Up to now, the main way of collecting ontologies has been through desk research following the strategies next described. However, as presented below, we also allow contributors to submit to the catalogue ontologies they are aware of.

3.1 *Review Literature*

Research literature and European project production (relevant projects as well as newly accepted projects) are the main sources for ontology search within the literature review.

3.2 *Analyze Standardization and Institutional Bodies*

One important source of ontological knowledge is normative standards and legislation. Standardization bodies, such as buildingSmart, ETSI (European Telecommunications Standards Institute), CEN (European Committee for Standardization), CENELEC (European Committee for Electrotechnical Standardization), W3C (World Wide Web Consortium), OASIS (Advancing Open Standards for the Information Society), OMG (Object Management Group), ISO (International Organization for Standardization), and OGC (Open Geospatial Consortium) usually provide standards that consist of data models with precise and useful descriptions of concepts (or even ontologies).

3.3 *Lookup Ontology Catalogues*

There are several catalogues that may be looked up for identifying relevant ontologies such as , LOV (Linked Open Vocabularies http://lov.okfn.org) (Vandenbussche and Vatant 2014), Watson (http://watson.kmi.open.ac.uk/) (D'Aquin and Motta 2011) or Swoogle (http://swoogle.umbc.edu/) (Finin et al. 2005).

3.4 *Dataset Investigation*

The investigation of available datasets will allow identifying the ontologies and vocabularies that they use. They may also provide links to other datasets as well as information about connections between these ontologies.

These datasets could be found in dataset catalogues such as Linked Data catalogues (for example, Datahub http://datahub.io/ and Reegle http://data.reegle.info/), open data catalogues (for example, the Open Data Index https://index.okfn.org/ and Open Government Data http://opengovernmentdata.org/data/), and web portals that contain datasets from a concrete organization or a domain (for example, the Buildings Performance Institute Europe (BPIE) http://www.buildingsdata.eu/datasearch and the University of Missouri web site https://library.missouri.edu/guides/data/inter-data/).

3.5 *Contact Stakeholders*

In order to collect ontologies from relevant stakeholders we follow a semi-automatic process that involves different people with different roles: a) **contributors**, who suggest ontologies to be included in the catalogue or even provide their descriptions (i.e., metadata) through an on-line form; b) **populators**, who include new ontologies into the catalogue by describing them through the on-line form; and c) **metadata curators**, who review, improve, and complete the meta-

data of the ontologies inserted by contributors and populators.

These roles and their interaction with the ontology collection process are illustrated in figure 1. As shown in such figure, the process consists of the following steps:

1. Contributors and populators provide ontology metadata through an on-line form. There is also an option for contributors to provide minimal information for ontologies by means of filling in a short on-line form; in this case, the metadata curators will be in charge of completing the ontology metadata.

2. The metadata is received by the curators, that is, the catalogue maintainers, who review, improve, and complete such data if needed. This step implies some manual evaluation of the collected metadata.

3. Once the metadata is curated, both an RDF (Brickley 2004) and an HTML representation of the catalogue information are generated (see section 7). During this process some evaluation tasks are carried out over the ontologies.

It should be noted that since the process contains a manual component (i.e., metadata curation) the catalogue is not immediately updated when a new ontology is introduced through the on-line form.

4 ONTOLOGY DESCRIPTION

As already mentioned, the main advantages of reusing existing ontologies for describing the data of the ontology catalogue are that the catalogue data will be more interoperable with existing data and that the time of developing the ontology for the catalogue decreases. For these reasons, a common set of metadata vocabularies has been reused to describe the ontologies that are included in the catalogue.

These metadata have been selected after analyzing two well-known ontologies that can be used to describe ontology metadata, namely, OMV (Ontology Metadata Vocabulary) (Hartmann et al. 2005) and VOAF (Vocabulary of a Friend http://purl.org/vocommons/voa).

One limitation of OMV is that it does not reuse terms already defined in other well-known ontologies. For this reason we follow the VOAF approach that consists on reusing terms already defined in other vocabularies and only add those strictly necessary. As a result, five vocabularies have been reused for describing the ontologies of the catalogue; their titles, prefixes and URIs are listed in table 1.

Figure 2 shows the ontology used to describe the ontologies included in the catalogue. As represented in such figure, the central class of the model is *voaf:Vocabulary*, that is used to represent ontologies. This class contains some attributes (or datatype properties) to represent the ontology title (*dc:title*), its description in natural language (*dc:description*), its creation date (*dc:issued*), its last modification date (*dc:modified*), its prefix (*vann:preferredNamespacePrefix*), and its namespace (*vann:preferredNamespaceUri*).

Individuals belonging to the class *voaf:Vocabulary* could be related to individuals belonging to other classes. This way, the ontology language in which an ontology is developed is stated through the relationship *omv:hasOntologyLanguage* between the classes *voaf:Vocabulary* and *omv:OntologyLanguage*; the syntax in which an ontology is available is represented by the relationship *omv:hasOntologySyntax* between the classes *voaf:Vocabulary* and *omv:OntologySyntax*; the domains covered by the ontology are indicated by means of the relationship *omv:hasOntologyDomain* between the classes *voaf:Vocabulary* and *omv:OntologyDomain*; the language in which the ontology is expressed is stated by the relationship *dc:language* between the classes *voaf:Vocabulary* and *dc:LinguisticSystem*; and the license of the ontology is indicated through the property *dc:license* between the classes *voaf:Vocabulary* and *cc:License*.

5 ONTOLOGY EVALUATION

One of the characteristics of the proposed catalogue is the evaluation component that provides an added value to the metadata gathered. Taking as inspiration the approach presented at (Poveda-Villalón et al. 2013), we follow the main guidelines for publishing data over the Web in order to provide quality indicators. First of all, we base the ontology evaluation features on the extremely well-known Linked Data principles and the Linked Open Data 5 Star rating system (http://www.w3.org/DesignIssues/LinkedData.html) defined by Tim Berners-Lee. More precisely, the rating system defines the following levels (taken literally from the source):

- *LOD1.* Available on the Web (whatever format) but with an open license, to be Open Data

- *LOD2.* Available as machine-readable structured data (e.g., excel instead of image scan of a table)

- *LOD3.* As (2) plus: non-proprietary format (e.g., CSV instead of excel)

- *LOD4.* All the above plus: use open standards from the W3C (RDF and SPARQL) to identify things, so that people can point at your stuff

- *LOD5.* All the above plus: link your data to other people's data to provide context

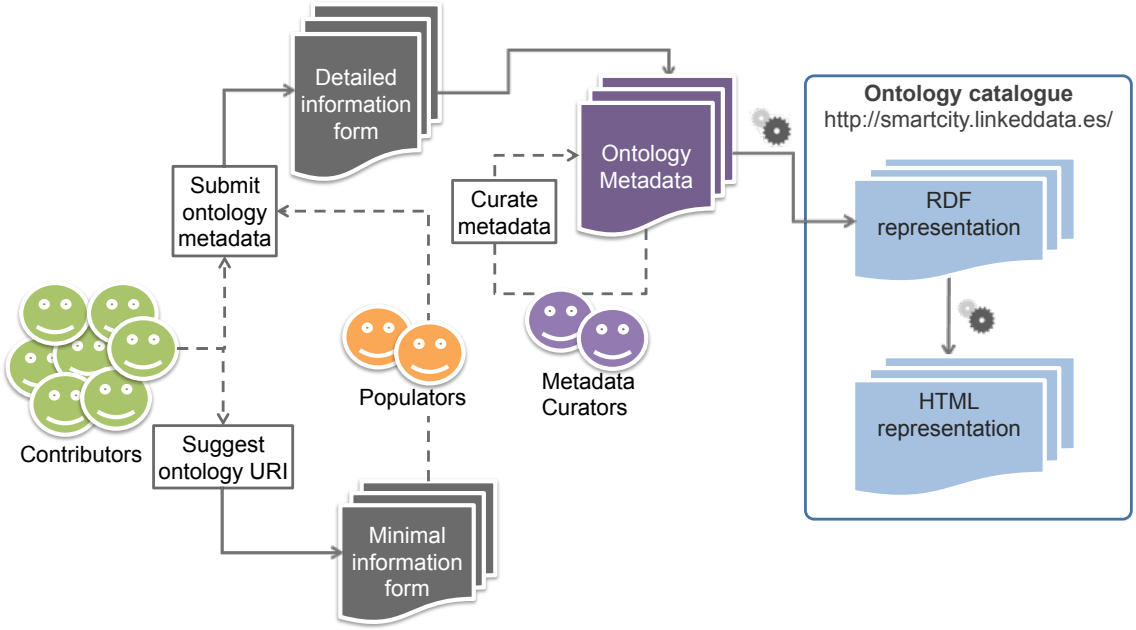

Figure 1: Proposed process to collect ontology metadata and generate the ontology catalogue.

Table 1: Vocabularies reused for describing the ontologies of the catalogue.

Vocabulary	Prefix	URI
Creative Commons Rights Expression Language	cc	http://creativecommons.org/ns
DCMI Metadata Terms	dc	http://purl.org/dc/terms/
Vocabulary of a Friend	voaf	http://purl.org/vocommons/voaf#
Ontology Metadata Vocabulary	omv	http://omv.ontoware.org/2005/05/ontology#
VANN: A vocabulary for annotating vocabulary descriptions	vann	http://purl.org/vocab/vann/

In addition, we also take into account the indicators provided by The Open Data Index (https://index.okfn.org/), which is used to measure how open is the data from a given country in a given domain. These indicators are:

- *ODI1*. Does the data exist?
- *ODI2*. Is it in digital form?
- *ODI3*. Is it publicly available?
- *ODI4*. Is it free of charge?
- *ODI5*. Is it online?
- *ODI6*. Is it machine-readable?
- *ODI7*. Is it available in bulk?
- *ODI8*. Is it openly licensed?
- *ODI9*. Is it up to date?

In first instance, the indicators for assessing ontologies taken into account in the catalogue are:

- Whether the ontology is available on the Web (whatever format). This indicator is related to LOD1 and ODI5. In this case the possible values for the indicator are "true" or "false" depending whether the ontology is available.
- Whether the ontology is available following the W3C standards (RDF-S, OWL or SKOS). This indicator is related to LOD4 and ODI6. In this case the possible values for the indicator are "true" if the ontology is provided in RDF-S, OWL or SKOS and "false" if the ontology is provided in any other formal language or it is not provided in any formal language.
- Whether the ontology is available under an open license. This indicator is related to LOD1 and ODI8. This indicator takes the value "true" if the license under which the ontology is published is an open license (http://licenses.opendefinition.org/#all-licenses) and "false" if the license is not open or if no license is specified.

6 CATALOGUE GENERATION

Once the model for describing ontology metadata is defined, the collected data from the on-line form is transformed into RDF according to such model.

The data gathered from the form is stored in a Comma-Separated-Value (CSV) file where each row contains the data related to a given ontology. For each ontology, an individual of *voaf:Vocabulary* is created, its attributes are filled in with the values introduced by the contributors or curators and, finally, the individual is linked to other individuals that represent on-

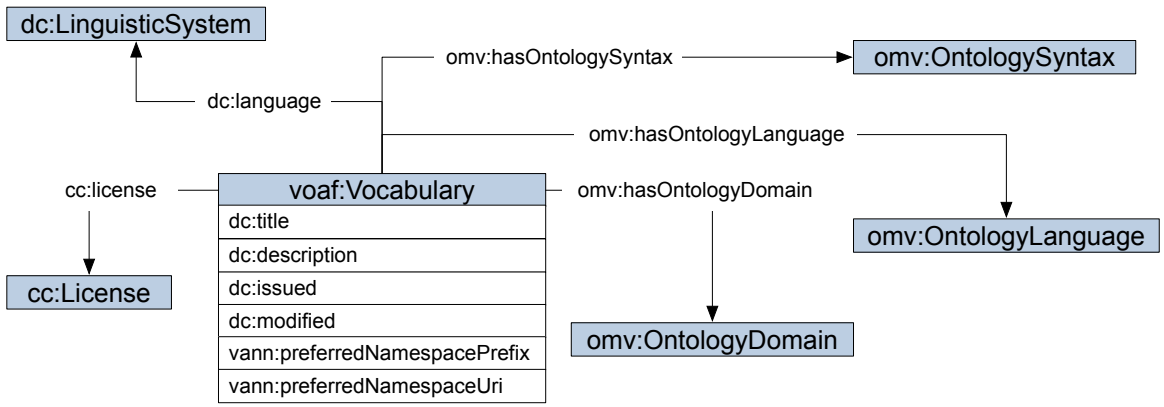

Figure 2: Ontology to represent ontology metadata.

tology syntaxes, ontology implementation languages, languages, licenses, and domains. An example of an ontology annotated following the ontology presented above is shown in figure 3.

It is worth noting that the generated RDF data is linked to existing datasets, following widely-used recommendations for publishing Linked Data (Bizer et al. 2009). In this case we use the DBpedia (http://dbpedia.org) and Lexvo (http://www.lexvo.org/) datasets in order to link to general domain and to linguistic entities, respectively. DBpedia may be considered the nucleus for the Web of Data since it contains information about a number of domains (e.g., geographic information, people, companies, online communities, films, music, books, among others) describing around 4.0 million entities for the English version. Lexvo has been selected to represent languages as it brings information over 7,000 language identifiers and ensures that these identifiers are dereferenceable and highly interconnected as well as externally linked to a variety of resources on the Web.

During this process there can appear different scenarios for attaching property values to a given individual of the class *voaf:Vocabulary*. The easiest case is when filling in the values for attributes because the value gathered from the CSV file is used as a Literal and directly linked to the vocabulary through the corresponding datatype property (for example *dc:title* in figure 2).

When the link between the vocabulary individual and the values to be attached is an object property (for example *omv:OntologyLanguage* in figure 2), it means that the target value takes the form of another individual, instead of a Literal. For these cases there are two possible ways of linking a given individual to other individuals. For the object properties *omv:hasOntologyLanguage*, *omv:hasOntologySyntax*, and *cc:License* there are sets of individuals pre-defined in the model because the possible values are an enumerated set. It should be noted that the current set of individuals might not cover all the cases; for example, for licenses, when a new license is included into the system a new individual for representing such a license is created. For the case of the object property *dc:language*, the Lexvo dataset is used in order to represent individuals of the class *dc:LinguisticSystem*.

In order to represent individuals form the classes *omv:OntologyLanguage*, *omv:OntologySyntax* and *cc:License* we give priority to URIs defined in official namespaces, that is, in namespaces controlled by the organism that created or maintains such concept or term. In this regard, we use URIs defined in the *cc* namespace to identify *cc* licenses (e.g., *http://creativecommons.org/licenses/by/3.0/*). If there is no official URI defining a given individual, we link to the corresponding DBpedia entity; for example, for representing the OWL ontology language we use *http://dbpedia.org/resource/Web_Ontology_Language*.

Finally, for representing the domains that a given ontology might cover there is no fixed set of possible individuals, that is, this field is a free text box in the on-line form where the contributor or curator could include any value or set of values. In order lo link the ontologies to the domains they are related to, we first try to find existing entities representing such domains. For doing so, ontology grounding techniques are used in order to determine links between the unrestricted terminology of users and resources of the Web of Data (particularly DBpedia), making easier the interoperability and later alignment among models (Lozano et al. 2012). If no entity from DBpedia is found, a new individual is created in a namespace under our control, as recommended in Linked Data development guidelines (Bizer et al. 2009). Among the advantages of linking ontology domains to DBpedia entities it should be noted: (a) the connection of the dataset being built with the Web of Data through a well connected dataset, DBpedia, and (b) the avoidance of duplicates due to different lexicalizations of the same concept.

As previously explained (see section 3) the collected data is generated both in machine-processable format (i.e., RDF data following the Linked Data principles) and in a human-readable documentation

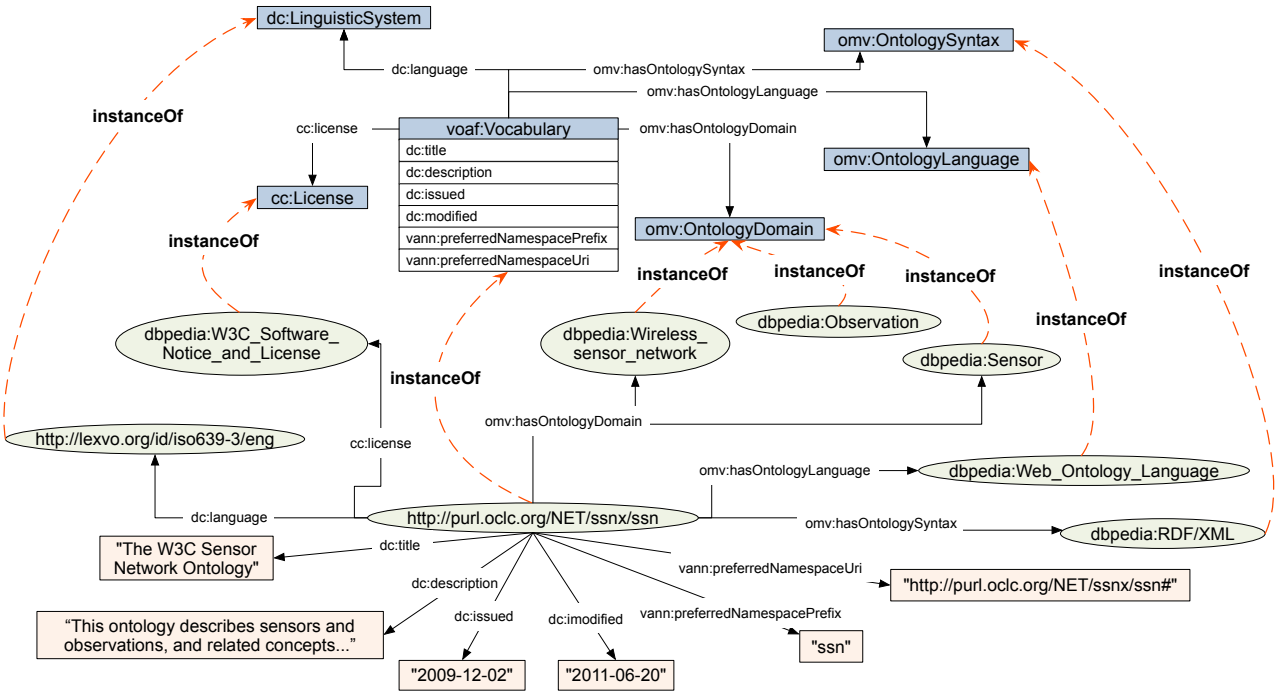

Figure 3: Example of ontology metadata representation in RDF.

(i.e., an HTML website that is automatically generated from the RDF).

7 WEB APPLICATION

The catalogue of ontologies about smart cities, energy and other related fields can be accessed through a web application available at http://smartcity.linkeddata.es/ontologies/.

As shown in figure 4 the catalogue allows visualizing metadata about the listed ontologies. For each ontology, the metadata are shown in the columns: “Open License”, “Ontology Language”, “Syntax”, “Domain” and “Natural Language”. The values shown in each cell of the table contain different information both represented by text and by color; for ontology metadata, colors have the following meaning: “plain information” for blue and “unknown” for grey. Furthermore, in addition to the color, each cell contains detailed information when available.

Apart from ontology metadata, the catalogue presents in the first three columns the quality indicators for the ontologies defined in section 5: “Online Availability’, “Open License”, and “Ontology Language”. For the quality indicators, colors have the following meaning: “success” for green, “warning” for orange, “danger” for red, and “unknown” for grey.

The values of the “Open License” and “Ontology Language” indicators are taken from the ontology metadata and the evaluation results are stated using color. For example, in the column “Open License” we can see that the ontologies “Units of Measure (OM)” and “The W3C Organization Ontology” are both published under an open license as the color of the cell is green, while detailed information about the licenses is also provided. More precisely, these licenses are “CC-BY 3.0” (Creative Commons Attribution 3.0 Unported) and “W3C” (W3C Software Notice and License) respectively, as shown in figure 4.

The “Online Availability” indicator represents whether the ontology is available in the Web in RDF and in HTML format. The evaluation of this indicator is performed on execution time when the catalogue is generated, that is, it is updated every time the catalogue is rebuilt.

Regarding this indicator, it is a good practice to provide information in different formats for a given resource in the Web (Bizer, Heath, & Berners-Lee 2009). In our case, the resource is the ontology itself and the different formats are its human-readable documentation (for example, an HTML page) and machine-readable information (for example, an RDF serialization). In addition, the mechanisms to provide both the HTML and RDF versions of the ontology should be compliant with the content negotiation recommendations (http://www.w3.org/TR/swbp-vocab-pub/). In order to automatically evaluate the indicator we first use Vapour (http://validator.linkeddata.org/vapour) to check whether the ontology URI provides RDF and HTML in a way compliant to content negotiation recommendations; if so, the color associated to the format is green. If for any format there is no correct content negotiation mechanism implemented, we next check whether any RDF and/or HTML resource is available even though the technical implementation is not compliant with content negotiation best practices; in that case, the format (RDF or HTML) is shown in orange. Finally, if one or both formats are not available through the URI, the format is represented in red.

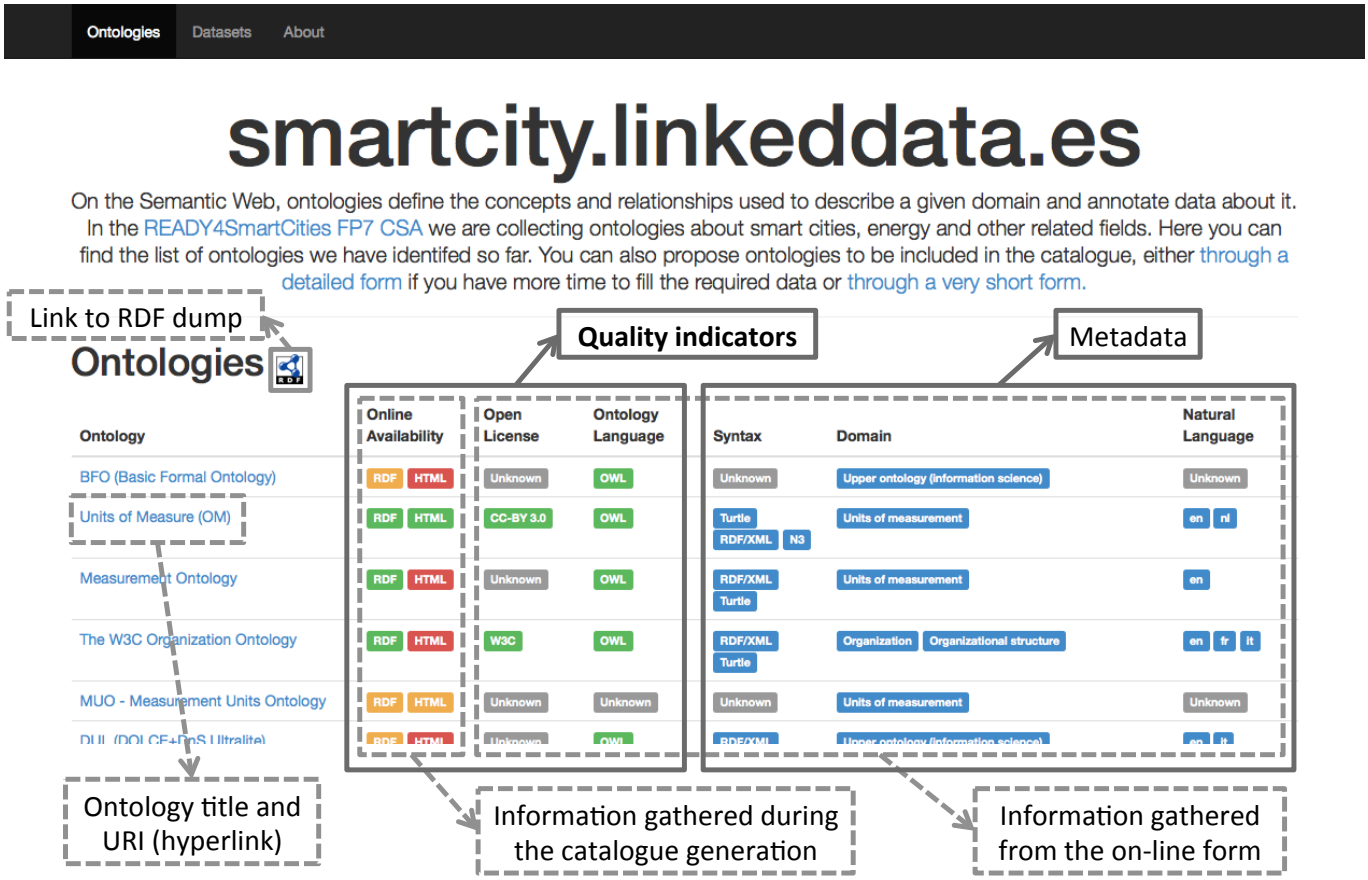

Figure 4: Screenshot of the ontology catalogue home page.

It should be noted that one ontology could provide one format according to content negotiation mechanisms and the other in a non-compliant way or not even provide it. Different combinations of these cases are shown in figure 4.

8 RELATED WORK

Due to the benefits of reusing existing ontologies, there have been several approaches to collect ontologies in order to provide ontology discovery services, such as ontology search engines or ontology registries (D'Aquin and Noy 2012). These registries and search engines could be categorized into two types, the general purpose ones containing general and domain ontologies; for example, LOV - Linked Open Vocabularies (Vandenbussche and Vatant 2014) http://lov.okfn.org, Swoogle http://swoogle.umbc.edu/, or Falcons http://ws.nju.edu.cn/falcons/, among others; and those focused in a given domain; for example, Bioportal (Whetzel et al. 2011), which is focused on biomedical ontologies.

The former ones are suitable for finding cross-domain ontologies. However, developers might spend too much time when looking for particular term in a specific knowledge area. For example, these systems would not be helpful to distinguish the term "point" in time, mathematic or location ontologies or might not even consider ontologies with a high degree of specialization on a given domain.

The main drawback about domain-specific registries is that they rarely can be shared across communities. So far, to the best of our knowledge, there is no ontology registry or index focused on smart cities and energy ontologies.

9 CONCLUSIONS AND FUTURE WORK

Along this paper we have shown the process followed to create an ontology catalogue for the smart cities and related domains (available at http://smartcity.linkeddata.es/ontologies/), in which information has been curated and that incorporates ontology evaluation features. This catalogue is the first approach within this community and it would be highly useful for new ontology developments or for describing and annotating existing ontologies.

As immediate future lines of work we plan to include faceted search as well as to provide a SPARQL endpoint so that users can query the RDF version of the catalogue. In addition, in order to provide a more detailed assessment (e.g., related to good modeling practices), the OWL ontologies available on the Web will be evaluated by means of external evaluation services such as OOPS! (OntOlogy Pitfall Scanner! - http://www.oeg-upm.net/oops) (Poveda-Villalón et al. 2012) which is an on-line application to identify pitfalls in ontologies.

Apart form the obvious benefits for the community as to ease the ontology search and reuse activities, the catalogue will also help to identify gaps within the

state of the art ontologies about smart cities and close domains. This identification would lead future ontology developments in this area.

This work has been carried out in the context of the READY4SmartCities European project. It is expected that it will include most of the relevant ontologies in the domains it focuses during the project lifetime. In addition, an stable version of the technology for generating the catalogue (both its RDF and HTML versions) is planned to be produced. Once the project is over, the sustainability of the software and the selected persistence mechanisms should be evaluated. Nonetheless, it is expected to have a low resource demand for maintaining the catalogue because, by then, the major part of the relevant ontologies will already be included in the catalogue, leaving little effort for collecting new ones.

ACKNOWLEDGMENTS

This work has been supported by the READY4SmartCities European project (608711). We are very grateful to READY4SmartCities partners for their feedback about the ontology catalogue.

REFERENCES

Bizer, C., T. Heath, & T. Berners-Lee (2009). Linked Data - the story so far. *International Journal on Semantic Web and Information Systems 5*, 1–22.

Brickley, D. (2004). RDF vocabulary description language 1.0: RDF schema. *http://www. w3. org/tr/rdf-schema/*.

D'Aquin, M. & E. Motta (2011). Watson, more than a semantic web search engine. *Semantic Web 2*(1), 55–63.

D'Aquin, M. & N. F. Noy (2012). Where to publish and find ontologies? a survey of ontology libraries. *Web Semantics: Science, Services and Agents on the World Wide Web 11*(0), 96 – 111.

Finin, T., L. Ding, R. Pan, A. Joshi, P. Kolari, A. Java, & Y. Peng (2005). Swoogle: Searching for knowledge on the semantic web. In *PROCEEDINGS OF THE NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE*, Volume 20, pp. 1682. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Hartmann, J., Raúl, Y. Sure, M. C. Suárez-Figueroa, P. Haase, A. Gómez-Pérez, & R. Studer (2005). Ontology metadata vocabulary and applications. In *On the Move to Meaningful Internet Systems 2005: OTM 2005 Workshops*, pp. 906–915. Springer.

Lozano, E., J. Gracia, O. Corcho, R. A. Noble, & A. Gómez-Pérez (2012, November). Problem-based learning supported by semantic techniques. *Interactive Learning Environments*.

Poveda-Villalón, M., M. C. Suárez-Figueroa, & A. Gómez-Pérez (2012). Validating ontologies with OOPS! In *Knowledge Engineering and Knowledge Management*, pp. 267–281. Springer.

Poveda-Villalón, M., B. Vatant, M. C. Suárez-Figueroa, & A. Gómez-Pérez (2013). Detecting good practices and pitfalls when publishing vocabularies on the web.

Vandenbussche, P.-Y. & B. Vatant (2014). Linked open vocabularies. *ERCIM news 96*, 21–22.

Whetzel, P. L., N. F. Noy, N. H. Shah, P. R. Alexander, C. Nyulas, T. Tudorache, & M. A. Musen (2011). Bioportal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. *Nucleic acids research 39*(suppl 2), W541–W545.